Korean Journal of Remote Sensing, Vol.36, No.5-3, 2020, pp.1053~1066
https://doi.org/10.7780/kjrs.2020.36.5.3.5

ISSN 1225-6161 ( Print )
ISSN 2287-9307 (Online)

Article

# 위성 자료와 수치모델 자료를 활용한 스태킹 앙상블 기반 $SO_2$ 지상농도 추정

최현영[1] · 강유진[1] · 임정호[2]† · 신민소[1] · 박서희[1] · 김상민[3]

# Monitoring Ground-level $SO_2$ Concentrations Based on a Stacking Ensemble Approach Using Satellite Data and Numerical Models

Hyunyoung Choi[1] · Yoojin Kang[1] · Jungho Im[2]† · Minso Shin[1] · Seohui Park[1] · Sang-Min Kim[3]

**Abstract:** Sulfur dioxide ($SO_2$) is primarily released through industrial, residential, and transportation activities, and creates secondary air pollutants through chemical reactions in the atmosphere. Long-term exposure to $SO_2$ can result in a negative effect on the human body causing respiratory or cardiovascular disease, which makes the effective and continuous monitoring of $SO_2$ crucial. In South Korea, $SO_2$ monitoring at ground stations has been performed, but this does not provide spatially continuous information of $SO_2$ concentrations. Thus, this research estimated spatially continuous ground-level $SO_2$ concentrations at 1 km resolution over South Korea through the synergistic use of satellite data and numerical models. A stacking ensemble approach, fusing multiple machine learning algorithms at two levels (i.e., base and meta), was adopted for ground-level $SO_2$ estimation using data from January 2015 to April 2019. Random forest and extreme gradient boosting were used as based models and multiple linear regression was adopted for the meta-model. The cross-validation results showed that the meta-model produced the improved performance by 25% compared to the base models, resulting in the correlation coefficient of 0.48 and root-mean-square-error of 0.0032 ppm. In addition, the temporal transferability of the approach was evaluated for one-year data which were not used in the model development. The spatial distribution of ground-level $SO_2$ concentrations based on the proposed model agreed with the general seasonality of $SO_2$ and the temporal patterns of emission sources.

**Key Words:** ground-level $SO_2$ concentrations, OMI, machine learning, stacking ensemble

Received October 5, 2020; Revised October 16, 2020; Accepted October 22, 2020; Published online October 27, 2020

[1] 울산과학기술원 도시환경공학과 석·박사과정생 (Combined MS/PhD Student, Department of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology)

[2] 울산과학기술원 도시환경공학과 정교수 (Professor, Department of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology)

[3] 국립환경과학원 기후대기연구부 환경위성센터 연구사 (Researcher, Environmental Satellite Center, Climate and Air Quality Research Department, National Institute of Environmental Research)

† Corresponding Author: Jungho Im (ersgis@unist.ac.kr)

**요약:** 이산화황($SO_2$)은 대기 중 화학 반응을 통해 2차 대기오염물질을 생성하는 전구체로, 주로 산업활동이나 주거 및 교통 활동 등을 통해 배출된다. 장기간 노출 시 호흡기 질환이나 심혈관 질환 등을 유발하여 인체 건강에 부정적인 영향을 미칠 수 있기 때문에 이에 대한 지속적인 모니터링이 필요하다. 우리나라에서는 $SO_2$에 대해 관측소 기반의 모니터링이 수행되고 있으나 이는 공간적으로 연속적인 정보를 제공하는 데에 한계가 있다. 따라서, 본 연구에서는 위성자료와 수치모델 자료를 융합하여 일별 13시를 타겟으로 하는 1 km의 고해상도로 공간적으로 연속적인 $SO_2$ 지상농도를 산출하였다. 2015년 1월부터 2019년 4월까지의 기간 동안 남한 지역에 대하여 스태킹 앙상블 기법을 이용하여 $SO_2$ 지상농도 추정 모델을 개발하였다. 스태킹 앙상블 기법이란 여러가지 기계학습 기법을 두 단계로 쌓는 방식으로 융합하여 단일 모델 대비 더 향상된 성능을 도출하는 방법이다. 본 연구에서는 베이스 모델로는 RF (Random Forest)와 XGB (eXtreme Gradient BOOSTing) 기법이, 메타 모델로는 MLR (Multiple Linear Regression) 기법이 사용되었다. 구축된 모델의 교차검증 결과 메타 모델은 상관계수(R) = 0.69와 root-mean-squared-error(RMSE) = 0.0032 ppm의 결과를 보였으며 이는 베이스 모델의 평균 대비 약 25% 향상된 안정성을 보였다. 또한 모델 구축에 사용되지 않은 기간에 대한 예측 검증을 수행하여 모델의 일반화 가능성을 평가하였다. 구축된 모델을 이용하여 남한 지역의 $SO_2$ 지상농도 공간분포를 분석한 결과 일반적인 계절성과 배출원의 변화를 잘 반영하는 패턴을 보임을 확인하였다.

## 1. 서론

대기 중의 이산화황($SO_2$)은 주로 산업활동으로 인한 화석연료 연소 및 발전소 가동 등의 인위적 발생원 혹은 화산 폭발과 같은 자연적 발생원에 의해 배출된다(Seo *et al.*, 2020; Pandey *et al.*, 2015). $SO_2$는 수 시간에서 수일까지의 비교적 짧은 체류시간을 가지는 대기오염물질로 광반응 혹은 촉매반응을 통해 삼산화황, 황산 혹은 황산염 등의 2차오염물질을 형성하는 전구체 역할을 한다(Kharol *et al.*, 2017). 이는 질소산화물($NO_x$)과 함께 산성비의 주요 원인 물질로서 토양 및 하천의 산성화에도 영향을 미치며 인체에 노출 시 점막을 자극하여 호흡기 질환과 심혈관질환을 유발하기도 한다. 급속한 산업화와 도시화로 인하여 최근 몇 년 동안 대기 오염원은 전 세계적으로 크게 증가하였으며 동아시아 지역은 $SO_2$ 농도가 높은 지역 중 하나로 널리 알려져 있다(Bauduin *et al.*, 2016). 특히 중국 북동부연안은 베이징 지역을 중심으로 산업 시설이 집중되어 있기 때문에 인공적 배출원이 많으며, 이는 편서풍을 타고 우리나라로 장거리 유입될 가능성이 높기 때문에 우리나라에서의 $SO_2$ 지상 농도를 지속적으로 모니터링하는 것은 중요한 실정이다(Lee, 2010).

우리나라에서는 관측소 기반의 $SO_2$ 지상농도 감시가 이루어지고 있지만 이는 점 기반의 관측 값으로 연속적인 공간 정보를 제공하지 못한다. 지상 관측소는 주로 도시 지역에 밀집되어 있는 불균형한 공간 분포를 가지기 때문에 대기오염 물질의 공간적 특성을 파악하는 데 한계가 있다(Zhang *et al.*, 2018). 이러한 한계점을 극복하기 위해 최근 넓은 지역에 대하여 연속적인 정보 제공이 가능한 위성 자료를 이용한 대기질 모니터링이 활발히 이루어지고 있다(Fernandes *et al.*, 2019; Park *et al.*, 2016). SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY (SCIAMACHY), Global Ozone Monitoring Experiment (GOME), Ozone Monitoring Instrument (OMI), TROPOspheric Monitoring Instrument (TROPOMI) 등의 위성들은 $SO_2$ 연직 컬럼 농도를 측정하여 제공한다. 위성으로 관측되는 농도는 연속적인 공간 분포를 가지지만 환경 및 인체에 직접적인 연관이 있는 지상농도 정보가 아닌 연직 컬럼 농도 형태로 제공되며 구름 하에서는 자료를 얻을 수 없기 때문에 수치모델 자료와의 융합을 통해 대기오염물질의 지상 농도를 산출하는 다양한 시도가 이루어지고 있다(Bauduin *et al.*, 2016; Zhang *et al.*, 2018; Zhang *et al.*, 2019).

$SO_2$ 지상농도 추정에 위성 자료를 사용할 경우 기본적으로 OMI를 포함한 위성 센서들이 대기 하층에서 산란과 흡수의 영향으로 약한 $SO_2$ 신호와 불확실성을 가진다는 점을 고려해야 한다(Kang *et al.*, 2019; Fioletov *et al.*, 2011). 월 평균 위성 기반 연직 컬럼 농도와 지상 관측 값은 0.49에 그치는 낮은 상관관계를 가지며(Zhang *et al.*, 2018), 이러한 비선형 관계를 해결하기 위해 $SO_2$ 지

상농도 추정에는 다양한 기계학습 방법이 사용되어 왔다(Li *et al.*, 2020; Zhang *et al.*, 2019). 하지만, 대부분의 선행연구는 단일 기법을 통해 낮은 시공간 해상도로 결과를 산출하여 $SO_2$의 짧은 체류시간(lifetime)에 의한 급격한 시공간적 변화를 모의하는 데에는 한계가 있다. 이에 본 연구에서는 다수의 기계학습 기법들을 융합하여 보다 일반화되고 우수한 성능을 가져오는 스태킹 앙상블(stacking ensemble) 기법을 적용하여 보다 높은 해상도에서의 시공간적 변화를 모의하고자 하였다. 따라서, 본 연구의 목표는 우리나라를 대상으로 지상관측 자료와 위성 및 수치모델 자료 등을 융합하여 1 km 고해상도 $SO_2$ 지상농도 추정 알고리즘을 개발하고, 제시된 모델의 시공간적 안정성을 검증하는 것이다.

## 2. 연구지역 및 연구자료

본 연구는 동경 124°-131°, 북위 33°-39°에 위치한 남한 지역에 대하여 수행되었으며(Fig. 1), 연구 기간은 2015년 1월부터 2019년 4월까지로 선정하였다. 이 지역은 전 세계적으로 대기오염물질의 농도가 높은 지역 중 하나로 알려진 동아시아 지역에 위치하고 있으며, 주로 인구가 밀집 되어있는 수도권이나 산업활동이 활발하게 이루어지는 공장 지대에서 고농도가 빈번하게 발생하는 양상을 보인다. 우리나라의 대기오염물질 관측소

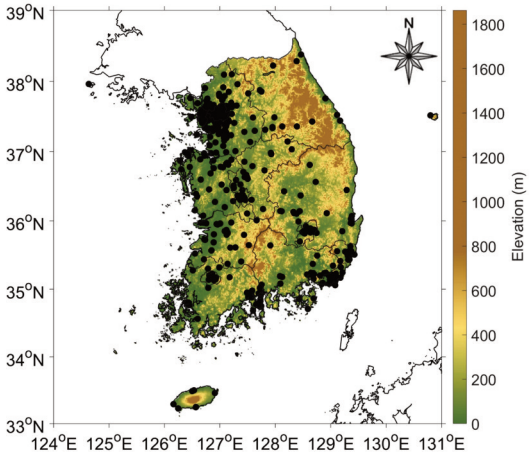


Fig. 1. Study area and distribution of $SO_2$ monitoring stations. Elevation (m) is used as a background image and the black points show the location of the ground monitoring stations.

는 도시 지역에 밀집되어 있고 산지나 교외 지역에는 드문 분포를 보이기 때문에, 연속적인 $SO_2$ 지상농도의 산출 및 분석을 위해 OMI 위성자료를 포함한 다양한 위성 산출물과 수치모델 기반 기상 및 배출량 자료, 관측소 기반 $SO_2$ 지상농도 자료 등을 함께 사용하였다.

남한의 $SO_2$ 지상 관측 자료를 본 연구에서 제시하는 모델의 타겟 변수로 사용하였으며, 에어코리아(Airkorea, https://www.airkorea.or.kr/web)에서 제공하는 확정자료를 사용하였다. 연구 기간 동안 관측소에서 1분 간격으로 측정되어 시간별 평균 값으로 제공되는 $SO_2$ 농도 값을 사용하였으며, OMI 위성의 local time (13:45)에 맞추어 13시와 14시 사이의 평균 농도인 13시 값을 사용하였다. 본 연구에서 사용된 $SO_2$ 지상농도 관측소의 개수는 총 409개로 Fig. 1과 같은 공간 분포를 가진다.

위성자료로는 Ozone Monitoring Instrument(OMI), Moderate Resolution Imaging Spectroradiometer(MODIS), 그리고 Global Precipitation Measurement(GPM) 세 가지 위성의 산출물을 사용하였다. OMI는 2004년 7월에 발사된 NASA(National Aeronautics Space Administration)의 EOS-Aura 위성에 탑재된 UV-VIS 분광계로 여러 미량기체 및 에어로졸의 특성을 관측한다. 극 궤도 위성인 OMI의 공간해상도는 0.25°×0.25°로 매일 약 13시 45분에 적도를 지나 24시간 동안 전 지구를 관측한다. OMI의 $NO_2$, $SO_2$, $O_3$, HCHO 연직 컬럼 농도 산출물인 OMNO2d, OMSO2e, OMDOAO3e, OMHCHOG를 사용하였다(NASA Goddard Earth Sciences Data and Information Services Center (GES DISC); https://mirador.gsfc.nasa.gov/). MODIS는 NASA에서 발사한 Terra 및 Aqua 위성에 탑재된 센서로 500 m 공간 해상도의 연간 토지 피복 산출물(MCD12Q1(Sulla-Menashe and Friedl, 2018); https://search.earthdata.nasa.gov/)을 제공한다. 본 산출물의 13×13개의 인접 픽셀을 이용한 이동 창(moving window)을 통하여 전체 토지 피복 유형 중 도시와 산림에 해당하는 픽셀 개수의 비율을 각각 계산하여 사용하였다. 추가적으로 강수가 $SO_2$에 미치는 영향을 고려하기 위해 GPM 위성의 30분 마다 제공되는 0.1° 해상도의 강수량 산출물(GPM_3IMERGHH)을 사용하였다. 강수는 침적 과정을 통해 다양한 대기오염물질을 제거하는 요인으로 잘 알려져 있으며, 특히 $SO_2$의 경우 물에 대한 용해도가 높아 다른 물질에 비해 더 큰 강수 세정

효과를 가진다(Yun, 2014; Westervelt *et al.*, 2017). 강수량 자료 역시 GES DISC에서 얻을 수 있으며 시간별 24시간 누적 강수량을 계산하여 사용하였다.

위성자료와 함께 수치모델 기반 기상학적 자료 및 배출량 자료를 사용하였다. 기상학적 자료는 기상청 기상자료개방포털에서 제공하는 지역예보모델(Regional Data Assimilation and Prediction System (RDAPS))을 사용하였고, 이의 공간해상도는 12 km이며 연직으로 약 80 km까지 70층으로 구성된다. RDAPS에서 하루 4번(00, 06, 12, 18시(KST)) 제공하는 분석장을 사용하였으며, 분석장을 제공하지 않는 시간에 대해서는 시간적 내삽을 통해 자료를 만들어준 뒤 입력변수로 사용하였다. 선행

Table 1. Input variables used to estimate ground-level $SO_2$ concentrations

| Category | Variables | Abbreviation | Spatial resolution | Temporal resolution |
|---|---|---|---|---|
| Satellite data | OMI $NO_2$ vertical column density | OMI $NO_2$ VCD | 0.25° | Daily |
| | OMI $SO_2$ vertical column density | OMI $SO_2$ VCD | | |
| | OMI $O_3$ vertical column density | OMI $O_3$ VCD | | |
| | OMI HCHO vertical column density | OMI HCHO VCD | | |
| | GPM Precipitation | Precipitation | 0.1° | 30 min |
| | MODIS Urban area ratio | LC_urban | 500 m | Yearly |
| | MODIS Forest area ratio | LC_forest | | |
| Meteorological data: RDAPS | Specific humidity | SPFH | 12 km | 4 / day |
| | Relative humidity | RH | | |
| | Visibility | VIS | | |
| | Dewpoint temperature | DPT | | |
| | Three-hour cumulative maximum wind speed | MAXGUST | | |
| | Friction velocity | FRICV | | |
| | Planetary boundary layer height | HPBL | | |
| | Surface pressure | PRES | | |
| | Surface temperature | SFCT | | |
| | Accumulated maximum wind speed (1,3,5 days) | StackMaxWS (1,3,5) | | |
| | Wind speed | WS | | |
| | Cosine value of wind direction | Wcos | | |
| | Sine value of wind direction | Wsin | | |
| Emission data: SMOKE | Methane ($CH_4$) | $CH_4$ | 9 km | Hourly |
| | Nitric oxide (NO) | NO | | |
| | Nitrogen dioxide ($NO_2$) | $NO_2$ | | |
| | Ammonia ($NH_3$) | $NH_3$ | | |
| | Formaldehyde (HCHO) | HCHO | | |
| | Carbon monoxide (CO) | CO | | |
| | Sulfur dioxide ($SO_2$) | $SO_2$ | | |
| | Primary organic aerosol | POA | | |
| | Primary nitrate | $PNO_3$ | | |
| | Primary sulfate | $PSO_4$ | | |
| | Other primary particulate matter ≤ 2.5 microns | PMFINE | | |
| Auxiliary data | Converted day of year | DOY | – | – |
| | Road density | RoadDens | 5 arc-minute (~8 km) | – |
| | Population density | PopDens | 30 arc-second (~1 km) | 5 years |

연구를 참고하여 $SO_2$ 농도와 연관성이 높은 RDAPS 기반 기상학적 변수 12개와 추가적으로 최대풍속을 1, 3, 5일 간격으로 누적시킨 자료를 사용하여 총 15개의 기상학적 변수가 사용되었다(Kim and Kim, 2014; Wang and Sun, 2019). 기상학적 변수와 함께 발전소, 항만 등에서의 국지적 배출의 영향을 고려하기 위하여 배출원 자료를 기반으로 하는 모델링 시스템인 SMOKE(Sparse Matrix Operator Kerner Emissions)에서 제공되는 11종의 오염물질 배출량 자료를 사용하였다. SMOKE는 국립환경과학원(National Institute of Environmental Research: NIER)에서 제공되고 있으며 1시간 간격으로 9 km의 공간 해상도의 자료가 제공된다.

위성 기반 및 수치모델 기반 자료 이외에 $SO_2$의 인위적 배출의 시공간적 패턴 정보를 고려하기 위해서 기타 보조 변수로 인구와 도로 밀도, 그리고 DOY(day of year)를 함께 사용하였다(Liu *et al.*, 2019; Wu *et al.*, 2020). 인구밀도 자료는 NASA Socioeconomic Data and Applications Center (SEDAC, https://sedac.ciesin.columbia.edu/)에서, 도로 밀도 자료는 GLOBIO (http://www.globio.info/download-grip-dataset)에서 다운로드 하였다. 본 연구에서는 위성자료와 수치모델 기반 자료, 그리고 기타 보조 자료들까지 총 36개의 변수를 사용하여 $SO_2$ 지상농도 산출 알고리즘을 개발하였으며, 사용된 모든 변수들은 Table 1에 약어 및 시공간 해상도와 함께 나타냈다.

## 3. 연구방법

본 연구에서 제시한 $SO_2$ 지상 농도 산출 알고리즘의 전반적인 흐름은 Fig. 2와 같다. 앞서 언급한 변수들은 모두 다른 공간 해상도를 가지기 때문에 1 km 격자에 맞추어 처리되었으며, 관측소 기반 $SO_2$ 지상농도의 경우 거리 가중 평균을 통해 한 픽셀에 하나의 농도 값이 되도록 계산하여 사용하였다. 이에 따라 본 연구에서 사용된 지상관측소의 개수는 409개이지만 1 km 격자 안에 두 개 이상의 관측소가 존재하는 경우 하나의 격자값으로 계산되기 때문에 궁극적으로 학습 시에 사용된 격자화된 관측소의 개수는 이보다 적은 398개이다. $SO_2$ 지상농도 산출을 위해 본 연구에서는 기계학습을 이용하였는데 모델의 안정성 확보를 위하여 두 단계로 이루

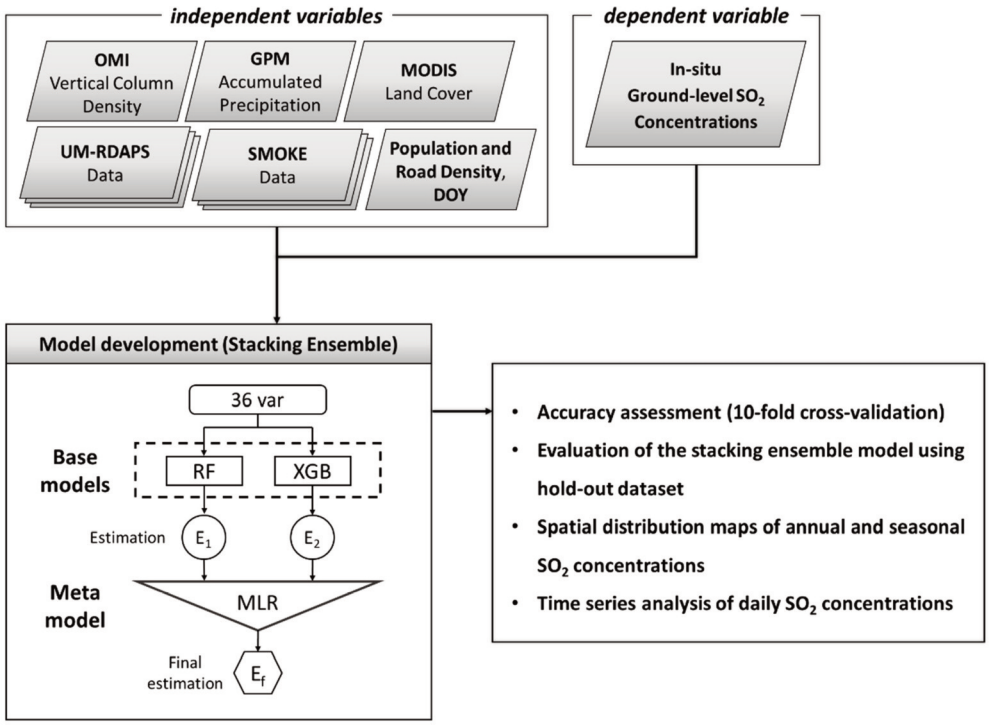


Fig. 2. Process flow diagram for estimation of ground-level $SO_2$ concentrations proposed in this study.

어진 스태킹 앙상블 기법을 적용하였다. 전체 연구기간 중 2018년을 제외한 나머지 기간의 자료는 모델 구축에 사용되었고, 구축된 모델을 2018년 자료에 적용하여 개발된 모델의 수행도를 검증하였다.

### 1) 기계학습: 스태킹 앙상블 기법

본 연구에서는 스태킹 앙상블 기법을 통해 $SO_2$ 지상농도를 추정하였으며, 이 기법은 최근 단일 학습자보다 일반화 및 견고성 측면에서 우수한 성능을 보이며 다양한 분류 및 회귀 문제에 사용되고 있다(Adhikari, 2015; Divina *et al.*, 2018; Herrera *et al.*, 2016; Ren *et al.*, 2016; Saini and Ghosh, 2017). 이는 서로 다른 다양한 모델을 층을 쌓듯이 조합하여 새로운 모델을 만드는 방식으로 베이스와 메타의 두 단계를 거치면서 각 모델들의 장점은 취하고 약점은 보완하면서 최종 모델의 성능을 향상시킨다(Xiao *et al.*, 2018).

스태킹 앙상블 기법에서는 베이스 모델의 예측 결과가 메타 모델의 입력변수로 사용되기 때문에 베이스 모델의 다양성과 각각의 성능이 최종 모델의 성능 개선에 중요한 역할을 한다(Feng *et al.*, 2020; Ren *et al.*, 2016; Scherbart and Nattkemper, 2008). 따라서 베이스 모델의 경우 RF(random forest), XGB(extreme gradient boosting), KNN(k-nearest neighbors), ANN(artificial neural network), SVR(support vector regression), 그리고 MLR(multiple linear regression)의 다양한 기계학습 기법들을 적용해본 후 그 중 가장 뛰어난 성능을 보이는 RF와 XGB를 사용하였다. 두 가지 기법을 제외한 다른 기법들은 상대적으로 떨어지는 성능을 보여 메타 모델의 입력변수로 사용할 경우 오히려 최종 모델의 성능을 저하시키는 결과를 보였다. RF와 XGB 두 가지 기법을 베이스 모델로 사용함으로써 모델의 variance를 감소시켜주는 bagging과 bias를 감소시켜주는 boosting의 효과를 동시에 얻을 수 있어 보다 일반화된 모델 구축이 가능하였다. 베이스 모델의 예측을 결합하기 위해 스태킹 앙상블의 두번째 단계인 메타 모델에서는 MLR 알고리즘이 적용되었으며 최종적으로 구축된 모델의 구조는 Fig. 2에서 나타내고 있는 바와 같다. MLR은 모델의 일반성과 안정성을 최대화하면서 베이스 모델의 신뢰도 값을 높이는 가장 간단한 방법으로, 훈련 자료에 과적합(overfit)될 가능성이 적다(Feng *et al.*, 2020). 따라서 MLR은 최근 여러 문헌에서 스태킹 앙상블 알고리즘의 메타 모델로 사용되고 있다(Chen *et al.*, 2019; Feng *et al.*, 2020). 최종적으로 본 연구에서는 메타 모델이 추정한 $SO_2$ 지상농도 값과 실제 현장 관측 농도 값을 비교함으로써 모델의 검증 및 시공간적 분석을 수행하였다.

### 2) 모델 평가

본 연구에서는 모델의 타당성을 평가하기 위해 두 가지 검증을 수행하였다: 1) 모델을 구축할 때 10-fold 교차검증(10-fold cross-validation)을 수행하였으며, 이는 전체 샘플을 랜덤하게 동일한 크기의 10개의 그룹으로 분할하여 하나의 그룹을 검증 데이터로 나머지 그룹은 학습데이터로 사용하여 검증하고 이와 같은 과정을 10번 반복하여 학습의 타당성을 검증하는 방법이다. 2) 모델 구축 후, 구축된 모델을 훈련 시에 사용되지 않은 2018년 자료에 적용하여 검증하는 홀드아웃 검증(hold-out validation)을 수행하였다.

개발된 모델의 성능을 평가하기 위해서는 기울기(slope), 상관계수(R), root-mean-square-error(RMSE), index of agreement(IA)와 variance of errors($E_v$)의 다섯 가지 널리 알려진 지표가 사용되었으며 각 지표는 다음과 같이 정의된다.

$$slope = \frac{\Delta x}{\Delta y} \tag{1}$$

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}} \tag{3}$$

$$IA = 1 - \left[\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(|x_i - \bar{y}| + |y_i - \bar{y}|)^2}\right] \tag{4}$$

$$E_v = var\left(\left|\frac{x_i - y_i}{y_i}\right|\right) \tag{5}$$

이 때 $x_i$와 $y_i$는 각각 $i$번째 샘플의 예측 값과 관측 값을 의미하며 $\bar{x}$와 $\bar{y}$는 예측 값과 관측 값의 평균을, $n$은 샘플의 개수를 나타낸다. IA는 0과 1 사이의 값으로 나타나는 모델 예측 오차의 표준화된 척도로써 1에 가까울수록 정확도가 높은 모델이라고 볼 수 있다(Li *et al.*, 2009). 안정성이 높은 모델은 외부 간섭의 영향을 덜 받으며 예측 결과의 신뢰성을 보장할 수 있기 때문에, 본 연구에서는 스태킹 앙상블 모델의 안정성을 평가하기

위하여 $E_v$를 사용하였다. $E_v$는 다양한 모델이 있을 때 각 모델의 안정성을 평가하는 간단하면서도 효과적인 지표이며(Li *et al.*, 2019; Sun and Li, 2020), $E_v$가 작을수록 안정성이 높다고 판단된다.

## 4. 연구결과 및 토의

### 1) 모델 검증

Table 2는 베이스 모델과 메타 모델의 10-fold 교차검증 결과이다. 베이스 모델의 평가 지표를 살펴보면 R과 $E_v$ 두 가지 지표에서 RF ($R$ = 0.69, $E_v$ = 0.3066)가 XGB ($R$ = 0.66, $E_v$ = 0.4134)에 비하여 더 우수한 성능을 보인다. 이는 RF가 실제 $SO_2$ 관측 값의 증감 경향을 더 잘 모의하고 있으며 비교적 안정적인 결과를 표출하고 있음을 의미한다. 반대로 기울기 측면에서는 XGB (slope = 0.47)가 RF (slope = 0.41)에 비하여 더 높은 값을 보인다. 이는 RF 기법이 전반적인 샘플에 대하여 과소 모의하는 결과를 보여주는 반면 XGB는 일부 샘플들을 높게 모의하여 전반적으로 기울기 값이 상승한 결과라고 할 수 있다. 즉, 두 개의 베이스 모델이 유사한 수준의 RMSE를 가지는 모델이라고 하더라도 RF는 대부분 음의 오차를 보이는 반면 XGB는 상대적으로 양의 오차를 보이는 샘플의 비율이 많다는 것을 의미한다. 이는 XGB의 $E_v$가 RF에 비해 더 높은 값을 가지는 것을 통해서도 확인 가능하다. RF와 XGB가 검증 결과에서 다른 경향을 보이는 것은 두 기법의 다른 성격에서 기인한다. 두 기법 모두 의사결정나무 기반의 기계학습 기법이지만 RF는 모든 의사 결정 나무의 값을 평균함으로써 안정적인 결과를 산출하고 variance를 줄이는 데 효과적인 bagging 기법인 반면, XGB는 순차적으로 훈련 자료의 bias를 줄이는 데에 초점을 맞춘 boosting 기법이다. 따라서 XGB는 평균을 통한 smoothing을 겪지 않는 방향으로 모델이 결정되기 때문에 RF에 비하여 다양한 결과 값을 산출하게 된다.

스태킹 앙상블 모델은 단일 모델들을 결합함으로써 일반화와 안정성을 장점으로 하는 기법으로 잘 알려져 있기 때문에(Zhai and Chen, 2018), 본 연구에서는 메타 모델의 안정성을 베이스 모델과 비교·검증하였다. 메타 모델의 평가 지표(slope = 0.48, R = 0.69, $E_v$ = 0.2684)는 RF의 장점인 높은 상관계수와 안정성, XGB의 장점인 높은 기울기를 모두 충족하는 결과를 보인다. 즉, 관측 값의 증감 경향을 안정적으로 모델링하는 동시에 과소 모의하던 경향을 상대적으로 개선시켰음을 의미한다. 베이스 모델과 메타 모델의 평가 지표들을 비교하였을 때, IA와 $E_v$ 두 가지 측면에서 어떤 베이스 모델에 대해서도 메타 모델이 우수한 성능을 보였다. 메타 모델의 $E_v$는 베이스 모델에 비하여 눈에 띄게 감소하였는데, 이는 베이스 모델의 평균 $E_v$ 값에 비하여 약 25% 향상된 결과이다. 즉, 스태킹 앙상블 모델이 두 베이스 모델의 장점을 잘 융합하여 전반적인 성능 향상을 이끌어 냈을 뿐 아니라 $SO_2$ 농도를 추정함에 있어 더 안정적인 모델임을 의미한다.

Table 3은 구축된 모델을 훈련 시에 사용되지 않은, 즉 시간적으로 분리되어 있는 2018년 자료에 적용하여 검증한 홀드아웃 검증 결과이다. 앞선 교차검증 결과와 비교하였을 때 베이스와 메타 모델 모두에서 slope와 RMSE 측면에서는 큰 차이가 나타나지 않았다. 그러나, R과 IA는 교차검증 결과에 비해 감소하였으며 $E_v$는 증가하였다. 특히 랜덤하게 구분된 10-fold 교차검증에 비하여 시간적으로 구분된 예측 검증에 대해서는 모델의 안정성에 대한 검증 결과($E_v$)가 다소 저하되는 경향을 보

Table 2. Accuracy statistics of 10-fold cross-validation for base and meta models

| Statistics | Base | | Meta |
|---|---|---|---|
| | RF | XGB | MLR |
| slope | 0.41 | 0.47 | 0.48 |
| R | 0.69 | 0.66 | 0.69 |
| RMSE (ppm) | 0.0032 | 0.0033 | 0.0032 |
| IA | 0.76 | 0.77 | 0.79 |
| $E_v$ | 0.3066 | 0.4134 | 0.2684 |

Table 3. Accuracy statistics of temporal hold-out validation for base and meta models

| Statistics | Base | | Meta |
|---|---|---|---|
| | RF | XGB | MLR |
| slope | 0.41 | 0.44 | 0.47 |
| R | 0.61 | 0.53 | 0.6 |
| RMSE (ppm) | 0.0031 | 0.0034 | 0.0031 |
| IA | 0.72 | 0.70 | 0.74 |
| $E_v$ | 0.6164 | 0.6992 | 0.5031 |

인다. 이는 교차검증의 경우에는 데이터셋이 랜덤하게 구분되기 때문에 같은 날짜의 샘플들이 일부는 훈련 자료로, 나머지는 검증 자료로 활용될 수 있기 때문이다. 만약 비슷한 위치의 샘플이 훈련 자료에 포함되어 있을 경우 교차검증 결과가 우수하게 나타날 수 있는 반면에 모델 훈련 기간에 포함되지 않은 날짜에 대한 예측 검증일 경우 이와 같은 경우가 발생하지 않는다. 이처럼 기계학습 기반의 모델이 훈련 기간에 포함되지 않는 기간에 대하여 모델의 성능이 감소하는 경향은 통계 및 기계학습 등 경험 모델에서 흔히 관찰되는 특징이다 (Huang *et al.*, 2018; S Park *et al.*, 2019; Shin *et al.*, 2020). 그럼에도 불구하고 메타 모델은 교차검증 결과와 비교하여 slope, R, RMSE와 IA 값이 크게 달라지지 않았으며, RF와 XGB 베이스 모델에 비해 우수한 예측 검증 결과를 보이고 있다. 따라서 예측 검증에 있어서도 스태킹 앙상블 모델이 가장 효과적인 모델임을 입증하였다.

### 2) $SO_2$의 시공간적 분포 분석

Fig. 3은 훈련에 사용되지 않은 2018년 한 해 동안의 연평균 $SO_2$ 지상농도의 공간적 분포를 나타낸 것이다. 위의 모델 검증 결과에서 훈련 기간에 포함되지 않은 기간에 대한 예측 검증 결과가 교차검증 결과에 비해 평가지표 상으로 다소 안정성이 떨어지는 것으로 나타났다. 따라서 개발된 모델의 공간적 분포 측면에서의 일반화 성능을 평가하기 위해 홀드아웃 검증 데이터셋을 통한 공간적 분포를 분석하였다. Fig. 3의 모델 예측 값과 관측 값을 비교하였을 때 고농도가 관찰되는 지역인 서울, 경기, 인천, 울산, 광양 등과 저농도가 관찰되는 지역인 강원도와 전라도 내륙의 공간적 분포가 잘 일치하는 것으로 나타났다. 모델링 결과의 공간적인 분포를 분석하기 위하여 베이스 모델인 RF와 XGB에서 제공하는 변수중요도를 고려하였다. 각 모델의 변수 중요도를 살펴보았을 때 두 모델에서 각각 변수중요도 상위 10 개 변수에서 공통적으로 나타나는 변수는 $PSO_4$, $SO_2$, POA, Wcos과 $PNO_3$로써 Wcos을 제외하고는 모두 SMOKE 모델 기반 배출량 변수이다. 즉, 이는 지상의 $SO_2$ 농도가 지역적인 오염물질 배출에 의한 영향을 크게 받는다는 것을 의미하며 산업 시설 및 공장 지대가 위치한 울산이나 광양에서 고농도가 나타나는 것을 뒷받침할 수 있다. 또한 인구가 밀집되어 산업 및 교통 활동이 활발하고 가정과 산업체에서 배출되는 오염물질 양이 많은 서울·경기 지역에서 나타나는 고농도 역시 뒷받침할 수 있다. 이는 한국 $SO_2$의 주 배출원이 인공적인 대기오염에 의한 것이라는 선행연구와 역시 일맥 상통한다(Khan *et al.*, 2017). 뿐만 아니라 선박에서는 상당한 양의 $SO_2$를 배출하는 것으로 알려져 있는데, 이러한 연안에서의 선박 활동은 해륙풍과 같은 국지풍의 영향으로 연안 지역의 농도에 영향을 줄 수 있기 때문에 우리나라의 대표적

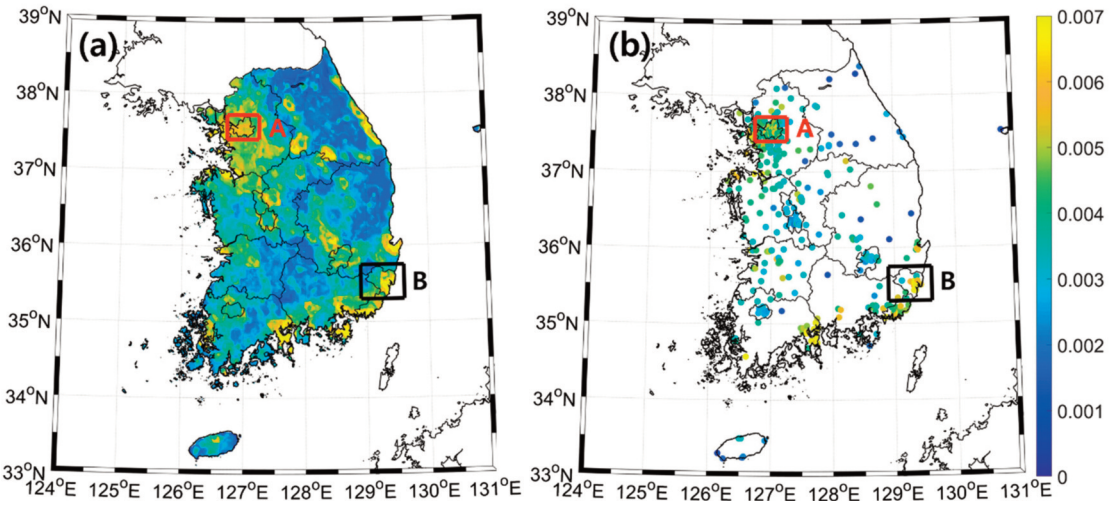


Fig. 3. Spatial distribution of annual $SO_2$ ground-level concentrations for 2018: (a) stacking ensemble model predictions and (b) observations. The unit of $SO_2$ concentration is part per million (ppm). 'A' box contains Seoul and 'B' box includes Ulsan.

인 항만 지역인 인천, 부산 및 울산에서의 고농도 관찰 원인 중 하나가 된다. 이처럼 모델 훈련에 포함되지 않은 기간에 대하여 개발된 모델을 적용하였을 때 관측소 값 및 선행연구와 일치하는 공간적 분포를 모의하는 것을 통해 모델의 일반화 성능을 확인하였다.

이와 같이 고농도 $SO_2$가 빈번히 발생하는 지역에 대하여 모델의 시공간적 모의 능력을 확인하기 위하여 Fig. 3에 나타난 서울 지역(A)과 울산 지역(B)에 대하여 모델 추정 값과 관측 값의 계절적 분포 변화를 분석하였다. Fig. 4와 Fig. 5은 각각 서울과 울산 지역의 계절별 $SO_2$ 지상농도 분포를 나타내며 모델 예측 값과 관측 값이 대체로 유사한 분포를 보이는 것을 확인할 수 있다. Fig. 4는 서울 지역에서 나타나는 뚜렷한 계절별 $SO_2$ 농도 변화를 잘 반영하고 있다. 서울에서는 여름과 가을에 낮은 평균 농도를, 봄과 겨울에 그 농도가 증가하는 일반적으로 널리 알려진 $SO_2$의 계절 변동 특성을 보인다. 겨울철에 농도가 증가하는 이유는 난방을 위한 연료 연소로 인한 증가가 일반적이며, 여름철 농도가 낮은 이유는 연료 연소가 줄어들고 강수가 증가하며 대류가 활발하게 일어나기 때문이다(Seung-Woo *et al.*, 2010). 스태킹 앙상블 모델 결과에서도 봄철과 겨울철에는 $SO_2$ 농도가 상승하는데, 이는 베이스 모델에서 공통적으로 중요한 변수로 나타난 $PNO_3$와 $SO_2$ 배출량이 계절적으로 봄과 겨울에 상승한 것을 원인으로 볼 수 있다. 실제 2018년 서울 지역의 $PNO_3$와 $SO_2$의 월별 배출량을 비교해보았을 때, $PNO_3$는 1, 2, 3, 4, 10, 12월에 높아지고, $SO_2$는 12, 1, 2월에 높아지는 경향을 보였다. 따라서 이러한 배출량의 증가가 서울과 경기지역의 $SO_2$ 지상 농도 증가로 이어졌다고 해석할 수 있다.

서울에서는 계절에 따른 뚜렷한 $SO_2$ 지상농도 증감이 관찰되는 반면, 울산에서는 서쪽과 동쪽으로 구분되어 공간적으로 변화가 다르게 나타난다(Fig. 5). 울산의 동쪽 지역에는 $SO_2$ 배출원인 석유화학단지와 온산국가산업단지가 위치하고 있으며, 이동오염원인 선박 운송이 활발하게 이루어진다(Oh *et al.*, 2016). 따라서, 내륙에 위치한 서쪽 지역에서는 계절에 따른 농도 변화가 거의

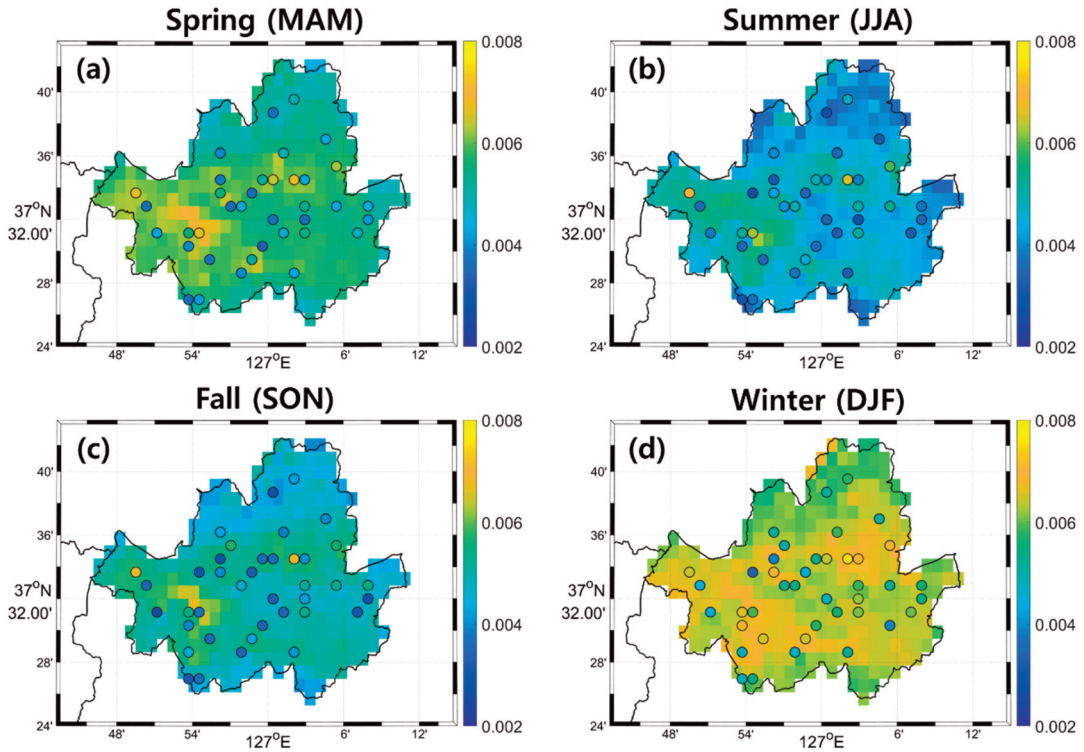


Fig. 4. Spatial distribution of ground-level $SO_2$ concentrations (ppm) by season ((a): spring (MAM), (b): summer (JJA), (c): fall (SON), (d): winter (DJF)) in Seoul. The background information represents the stacking ensemble model prediction and the dots represent the station observations.

없지만 해양과 맞닿아 있는 동쪽 지방에서는 계절에 따른 농도 변화가 매우 크게 나타난다. 한 가지 더 눈에 띄는 점은 $SO_2$의 일반적인 계절적 변동과 달리 울산의 동쪽 지역에서는 겨울철에 오히려 농도가 감소하는 경향을 보인다는 것이다. 봄철에 농도가 점점 상승하다가 여름철에 고농도가 내륙까지 넓게 유입되는 것이 관찰되며, 가을철과 겨울철이 되면서 점차 해안지역에서만 고농도가 관찰된다. 이는 울산항에서의 활발한 선박 운송과 계절풍 및 해륙풍의 영향으로 해석할 수 있다. 우리나라는 종관규모에서 여름철에는 남동풍이, 겨울철에는 북서풍이 불며, 국지규모에서는 낮 시간에는 해풍이, 새벽 시간에는 육풍이 분다. 따라서 본 연구의 타겟 시간인 13:00 KST는 낮 시간으로 해풍이 우세하다. 이러한 국지풍이 종관규모의 계절풍과 결합하여 여름철에는 연안에서 내륙으로 강한 바람이 불어 선박 배출원이나 공장 지대에서 발생한 $SO_2$를 내륙까지 운반하여 고농도가 멀리 유입되도록 한다. 실제로 해풍 유입 시 대기오염물질이 내륙으로 운반되어 해안에서 약 20 km 떨어진 지점까지도 영향을 미칠 수 있으며, 따라서 연안도시지역의 국지풍이 오염물질의 분포에 미치는 영향은 여러 선행연구에서 입증된 바 있다(Oh *et al.*, 2004; Lee *et al.*, 2013; Shang *et al.*, 2019). 반대로 겨울철에는 내륙에서 바다 쪽으로 강한 바람이 불어 선박이나 공장 지대에서 배출된 대기오염물질이 내륙으로 유입되는 것을 막아준다(Lee *et al.*, 2014). 이처럼 서울과 울산 지역의 예를 통해 본 모델이 다양한 배출량 및 기상학적 변수들을 함께 고려함으로써 $SO_2$ 지상농도의 시공간적 변동을 잘 모의하고 있음을 시사한다.

계절별 공간적 분포 분석과 더불어 본 모델의 일별 모의 능력을 검증하기 위하여 대표적으로 계절에 따른 변화가 잘 드러났던 울산 지역에 위치한 관측소에서의 지상 관측 값과 모델 예측 값의 시계열 분포를 확인하였다. 분석에 사용된 관측소는 울산광역시 북구 효문동에 위치하고 있으며(Fig. 5), 본 관측소 주변에는 현대자

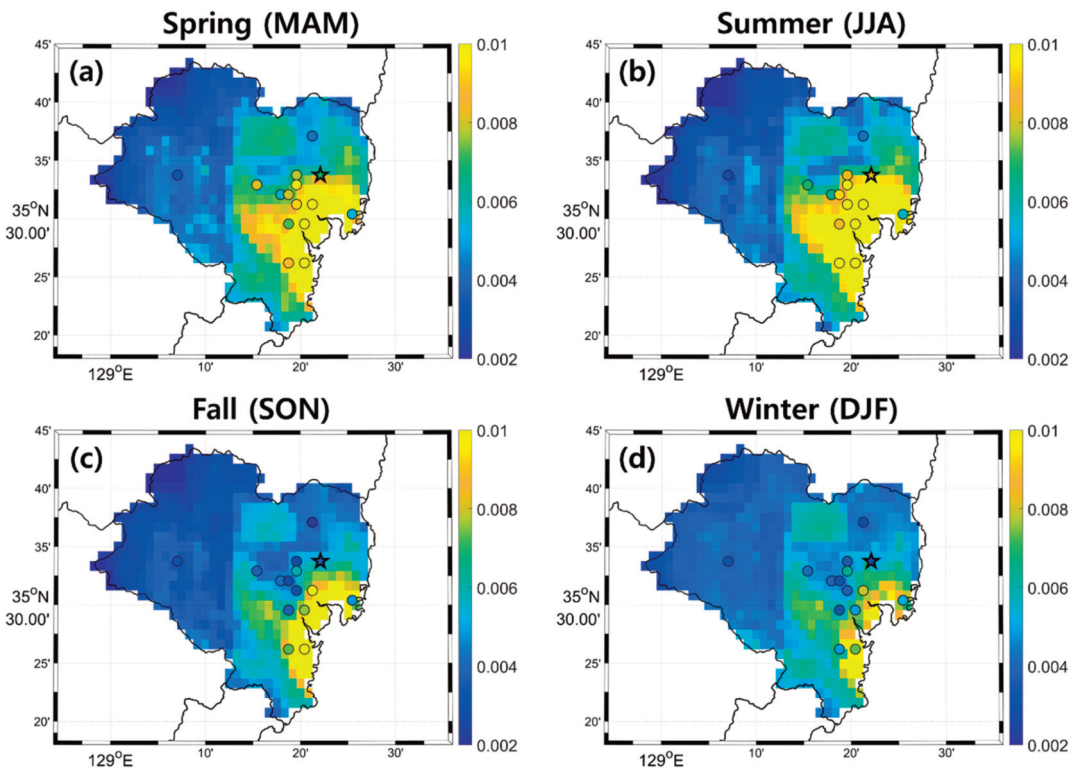


Fig. 5. Spatial distribution of ground-level $SO_2$ concentrations (ppm) by season ((a): spring (MAM), (b): summer (JJA), (c): fall (SON), (d): winter (DJF)) in Ulsan. The background information represents the stacking ensemble model prediction and the dots represent the station observations. The station marked with an asterisk was used for time series analysis (Hyomun-dong station).

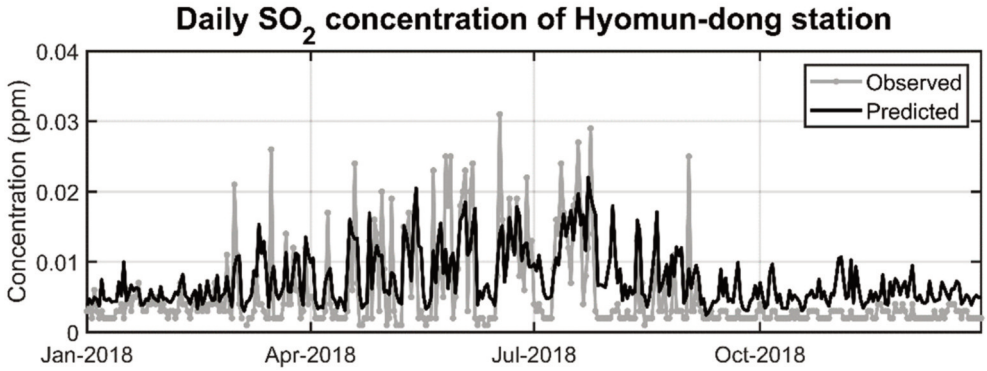


Fig. 6. Time series of daily observed(grey) and predicted(black) $SO_2$ concentration in Hyomun-dong station through 2018.

동차시트공장, 삼호철강 등 다양한 산업 시설들이 위치하고 있다. 공간적 분포 분석과 마찬가지로 시간적 분석 역시 훈련에 사용되지 않은 홀드아웃 검증 데이터셋(2018년)을 통하여 개발된 모델의 일반화 성능을 평가하였다. Fig. 6은 2018년 한 해 동안 효문동 관측소에서의 일별 $SO_2$ 관측 값과 예측 값의 시계열 그래프로, 여름철에는 본 관측소가 위치한 지역까지 고농도가 유입되며 겨울철에는 오히려 농도가 감소하는 울산 동부 지역의 일반적인 경향을 띈다. Fig. 6에서 본 모델은 저농도 사례에 대해서는 비교적 잘 모의하고 있으나 봄철과 여름철 고농도 사례들에 대해서는 실제 관측 농도보다 낮은 농도로 예측하고 있다. 고농도 사례에서는 음의 편차가 존재하였으며, 이는 Table 3에서 스태킹 앙상블 모델이 낮은 기울기 값을 가지는 원인으로 작용한다. 본 시계열 그래프 상에 나타난 관측 값과 예측 값 사이의 상관계수는 0.65로, 홀드아웃 검증 결과의 상관계수인 0.6과 비슷한 수준을 보이므로 훈련에 사용되지 않은 기간에 대한 모델의 일반성이 확인되었다. 따라서, 추후 훈련 자료에서의 고농도 샘플에 대한 보완이나 편차보정 등을 통하여 모델의 성능을 향상시킬 수 있을 것으로 기대된다.

## 5. 결론

본 연구에서는 2015년 1월부터 2019년 4월까지에 대해 위성자료와 수치모델 자료, 기타 보조자료를 융합하여 기계학습을 적용해 남한 지역의 $SO_2$ 지상 농도를 추정하였다. RF와 XGB를 베이스 모델로, MLR을 메타 모델로 한 스태킹 앙상블 모델을 구축하였으며 두 단계를 거쳐 다양한 기법들을 융합함으로써 단일 모델에 비해 향상된 결과를 얻었다. 이는 slope, R, IA가 증가하고 RMSE 와 $E_v$가 감소하는 것을 통해 검증되었다.

추정된 $SO_2$ 지상농도 값과 관측소 농도 값의 시공간적 분포는 일치하는 양상을 보였으며 국지적 배출의 영향을 많이 받는 도시 및 산업 지역에서 나타나는 고농도 분포 역시 잘 모의하고 있다. RF와 XGB 모델의 변수 중요도에 따르면 $PSO_4$, $SO_2$, $PNO_3$, POA와 같은 배출량 자료가 $SO_2$ 지상농도 추정에 중요한 역할을 하는 것으로 나타났다. 이는 $SO_2$ 지상농도가 지역적인 오염물질 배출의 영향을 크게 받는 것을 의미하며, 주거 및 산업활동 등이 발달한 지역에서의 고농도 발생을 대변해준다. 또한 계절별 배출원의 변화뿐 아니라 종관 및 국지규모의 기상 상황 역시 $SO_2$ 지상농도의 공간적 분포에 영향을 미치는 것으로 나타났다. 본 연구에서 제안한 모델은 높은 시공간 해상도로 연속적인 관측을 가능하게 하여 대기질 모니터링 측면에서 큰 기여를 할 것으로 사료된다.

하지만 스태킹 앙상블은 관측 값을 타겟으로 훈련하여 결과를 도출해내는 통계적 모델로서 훈련 샘플이 부족한 지역에서는 성능이 떨어진다는 한계점이 있다. 따라서 관측소가 없는 지역에서 훈련 샘플들과 매우 다른 특성을 가지는 경우 이러한 사례를 학습하기 어려워 그 지역에 대한 정확한 추정이 어려운 단점이 있다. 특히 $SO_2$의 경우 고농도 발생 비율이 매우 적어 고농도 사례에 대한 모의가 여전히 한계점으로 남아있으며, 본 모

델의 검증 결과 기울기 측면에서 저 추정하는 경향 역시 편향된 농도 분포에 의해 발생한 것으로 보인다. 향후 고농도 샘플의 비율을 임의로 늘려주어 샘플 분포의 불균형을 해결해주는 오버샘플링(oversampling) 기법을 적용하거나 혹은 저 추정된 모델의 예측 값을 편차보정(bias correction)을 통해 보완해줄 경우 모델 개선 가능성이 기대된다. 공간적 측면에서는 OMI의 후속 위성이며 보다 높은 공간 해상도 (5.5 km × 3.5 km)의 연직 컬럼 농도 산출물을 제공하는 TROPOMI 위성을 활용할 경우 개선된 성능을 보일 것으로 판단된다. 또한 2020년 2월에 발사된 우리나라의 정지궤도 환경위성(Geostationary Environment Monitoring Spectrometer, GEMS) 역시 본 연구에 사용된 4 종의 미량기체 연직 컬럼 농도를 제공하기 때문에 품질검증 과정을 거친 후 본 모델에 적용하여 활용 가능할 것으로 보인다. OMI에 비해 더 높은 공간해상도(7 km × 8 km)와 일 8회의 관측횟수로 우리나라 및 동아시아 지역의 대기환경을 연속적으로 감시할 수 있어 향후 $SO_2$ 지상농도에 대한 보다 정확한 정보 제공이 가능해질 것으로 기대된다.

## 사사

본 논문은 환경부의 재원으로 국립환경과학원의 지원을 받아 수행하였고(NIER-SP2020-01-02-008), 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음(IITP-2020-2018-0-01424).

## References

Adhikari, R, 2015. A neural network based linear ensemble framework for time series forecasting, *Neurocomputing*, 157: 231-242.

Bauduin, S., L. Clarisse, J. Hadji-Lazaro, N. Theys, C. Clerbaux, and P.-F. Coheur, 2016. Retrieval of near-surface sulfur dioxide ($SO_2$) concentrations at a global scale using IASI satellite observations.

Chen, J., J. Yin, L. Zang, T. Zhang, and M. Zhao, 2019. Stacking machine learning model for estimating hourly PM2. 5 in China based on Himawari 8 aerosol optical depth data, *Science of The Total Environment*, 697: 134021.

Divina, F., A. Gilson, F. Goméz-Vela, M. García Torres, and J. F. Torres, 2018. Stacking ensemble learning for short-term electricity consumption forecasting, *Energies*, 11(4): 949.

Feng, L., Y. Li, Y. Wang, and Q. Du, 2020. Estimating hourly and continuous ground-level PM2.5 concentrations using an ensemble learning algorithm: The ST-stacking model, *Atmospheric Environment*, 223: 117242.

Fernandes, A., M. Riffler, J. Ferreira, S. Wunderle, C. Borrego, and O. Tchepel, 2019. Spatial analysis of aerosol optical depth obtained by air quality modelling and SEVIRI satellite observations over Portugal, *Atmospheric Pollution Research*, 10(1): 234-243.

Fioletov, V., C. McLinden, N. Krotkov, M. Moran, and K. Yang, 2011. Estimation of $SO_2$ emissions using OMI retrievals, *Geophysical Research Letters*, 38(21).

Herrera, F., F. Charte, A. J. Rivera, and M. J. Del Jesus, 2016. Multilabel classification, in *Multilabel Classification*, edited, pp. 17-31, Springer.

Huang, K., Q. Xiao, X. Meng, G. Geng, Y. Wang, A. Lyapustin, D. Gu, and Y. Liu, 2018. Predicting monthly high-resolution PM2. 5 concentrations with random forest model in the North China Plain, *Environmental Pollution*, 242: 675-683.

Kang, H., J. Park, J. Yang, W. Choi, D. Kim, and H. Lee, 2019. Uncertainties of $SO_2$ Vertical Column Density Retrieval from Ground-based Hyperspectral UV Sensor Based on Direct Sun Measurement Geometry, *Korean Journal of Remote Sensing*, 35(2): 289-298.

Khan, A., K.-H. Kim, J. E. Szulejko, R. J. Brown, E.-C. Jeon, J.-M. Oh, Y. S. Shin, and A. A. Adelodun, 2017. Long-term trends in airborne $SO_2$ in an

air quality monitoring station in Seoul, Korea, from 1987 to 2013, *Journal of the Air & Waste Management Association*, 67(8): 923-932.

Kharol, S. K., C. A. McLinden, C. E. Sioris, M. W. Shephard, V. Fioletov, A. van Donkelaar, P. Sajeev, and R. V. Martin, 2017. OMI satellite observations of decadal changes in ground-level sulfur dioxide over North America, *Atmospheric Chemistry and Physics*, 17(9): 5921.

Kim, B.-W., and K.-H. Kim, 2014. The Analysis of Time Series of $SO_2$ Concentration and the Control Factor in An Urban Area of Yongsan-gu, Seoul, *Journal of the Korean Earth Science Society*, 35(7): 543-553.

Lee, H.-D., G.-H. Lee, I.-D. Kim, J.-S. Kang, and K.-J. Oh, 2013. The influences of concentration distribution and movement of air pollutants by sea breeze and mist around Onsan industrial complex, *Clean Technology*, 19(2): 95-104.

Lee, H. D., J. W. Yoo, M. K. Kang, J. S. Kang, J. H. Jung, and K. J. Oh, 2014. Evaluation of concentrations and source contribution of $PM_{10}$ and $SO_2$ emitted from industrial complexes in Ulsan, Korea: Interfacing of the WRF-CALPUFF modeling tools, *Atmospheric Pollution Research*, 5(4): 664-676.

Li, H., F. Faruque, W. Williams, M. Al-Hamdan, J. Luvall, W. Crosson, D. Rickman, and A. Limaye, 2009. Optimal temporal scale for the correlation of AOD and ground measurements of $PM_{2.5}$ in a real-time air quality estimation system, *Atmospheric Environment*, 43(28): 4303-4310.

Li, H., J. Wang, R. Li, and H. Lu, 2019. Novel analysis-forecast system based on multi-objective optimization for air quality index, *Journal of Cleaner Production*, 208: 1365-1383.

Li, R., L. Cui, J. Liang, Y. Zhao, Z. Zhang, and H. Fu, 2020. Estimating historical $SO_2$ level across the whole China during 1973-2014 using random forest model, *Chemosphere*, 247: 125839.

Liu, Q., S. Wang, W. Zhang, J. Li, and G. Dong, 2019. The effect of natural and anthropogenic factors on $PM_{2.5}$: Empirical evidence from Chinese cities with different income levels, *Science of the Total Environment*, 653: 157-167.

Oh, I., J.-H. Bang, S. Kim, E. Kim, M.-K. Hwang, and Y. Kim, 2016. Spatial distribution of air pollution in the Ulsan metropolitan region, *Journal of Korean Society for Atmospheric Environment*, 32(4): 394-407.

Oh, I., Y. Kim, and M. Hwang, 2004. Effects of late sea-breeze on ozone distributions in the coastal urban area, *J. Kor. Soc. Atmos. Environ*, 20: 345-360.

Pandey, A. K., R. P. Kumar, and K. Kumar, 2015. Satellite and ground-based seasonal variability of $NO_2$ and $SO_2$ over New Delhi, India, paper presented at Remote Sensing of Clouds and the Atmosphere XX, International Society for Optics and Photonics.

Park, J., J. Ryu, D. Kim, J. Yeo, and H. Lee, 2016. Long-range transport of $SO_2$ from continental Asia to northeast Asia and the northwest Pacific ocean: flow rate estimation using OMI data, surface in situ data, and the HYSPLIT model, *Atmosphere*, 7(4): 53.

Park, S., M. Shin, J. Im, C.-K. Song, M. Choi, J. Kim, S. Lee, R. Park, J. Kim, and D.-W. Lee, 2019. Estimation of ground-level particulate matter concentrations through the synergistic use of satellite observations and process-based models over South Korea.

Ren, Y., L. Zhang, and P. N. Suganthan, 2016. Ensemble classification and regression-recent developments, applications and future directions, *IEEE Computational Intelligence Magazine*, 11(1): 41-53.

Saini, R., and S. Ghosh, 2017. Ensemble classifiers in remote sensing, *Proc. of A review, paper presented at 2017 International Conference on Computing, Communication and Automation (ICCCA), IEEE*, Greater Noida, India, May. 5-6, pp. 1148-1152.

Scherbart, A., and T. W. Nattkemper, 2008. The diversity of regression ensembles combining bagging and random subspace method, paper presented at *International Conference on Neural Information Processing*, Springer, Berlin, Heidelberg, Germany, pp. 911-918.

Seo, J., J. Yoon, G.-H. Choo, D.-r. Kim, and D.-W. Lee, 2020. Long-term Trend Analysis of $NO_x$ and $SO_x$ over in East Asia Using OMI Satellite Data and National Emission Inventories (2005-2015), *Korean Journal of Remote Sensing*, 36(2-1): 121-137 (in Korean with abstract).

Seung-Woo, L., C.-H. Lee, D.-H. Ji, and H.-J. Youn, 2010. Temporal and spatial distribution of ambient sulfur dioxide concentration in forest areas, Korea, *Korean Journal of Soil Science and Fertilizer*, 43(6): 1035-1039 (in Korean with abstract).

Shang, F., D. Chen, X. Guo, J. Lang, Y. Zhou, Y. Li, and X. Fu, 2019. Impact of Sea Breeze Circulation on the Transport of Ship Emissions in Tangshan Port, China, *Atmosphere*, 10(11): 723.

Shin, M., Y. Kang, S. Park, J. Im, C. Yoo, and L. J. Quackenbush, 2020. Estimating ground-level particulate matter concentrations using satellite-based data: a review, *GIScience & Remote Sensing*, 57(2): 174-189.

Sulla-Menashe, D., and M. A. Friedl, 2018. User guide to collection 6 MODIS land cover (MCD12Q1 and MCD12C1) product, *USGS: Reston*, VA, USA, 1-18.

Sun, W., and Z. Li, 2020. Hourly $PM_{2.5}$ concentration forecasting based on feature extraction and stacking-driven ensemble model for the winter of the Beijing-Tianjin-Hebei area, *Atmospheric Pollution Research*, Elsevier, Amsterdam, NL, Jun. 6, vol. 11, pp. 110-121.

Wang, X., and W. Sun, 2019. Meteorological parameters and gaseous pollutant concentrations as predictors of daily continuous $PM_{2.5}$ concentrations using deep neural network in Beijing-Tianjin-Hebei, China, *Atmospheric Environment*, 211: 128-137.

Westervelt, D., A. Conley, A. Fiore, J. F. Lamarque, D. Shindell, M. Previdi, G. Faluvegi, G. Correa, and L. Horowitz, 2017. Multimodel precipitation responses to removal of US sulfur dioxide emissions, *Journal of Geophysical Research: Atmospheres*, 122(9): 5024-5038.

Wu, Y., R. Li, L. Cui, Y. Meng, H. Cheng, and H. Fu, 2020. The high-resolution estimation of sulfur dioxide ($SO_2$) concentration, health effect and monetary costs in Beijing, *Chemosphere*, 241: 125031.

Xiao, Y., J. Wu, Z. Lin, and X. Zhao, 2018. A deep learning-based multi-model ensemble method for cancer prediction, *Computer methods and programs in biomedicine*, 153: 1-9.

Yun, S.-Y., 2014. Seasonal washout effect of precipitation on major air pollutants ($O_3$, CO, $NO_2$, $SO_2$, $PM_{10}$) during summer and winter, Unpublished master's thesis, Ewha Womans University, Seoul, Korea.

Zhai, B., and J. Chen, 2018. Development of a stacked ensemble model for forecasting and analyzing daily average $PM_{2.5}$ concentrations in Beijing, China, *Science of The Total Environment*, 635: 644-658.

Zhang, D., Y. Zhou, L. Zheng, R. Shi, and M. Chen, 2018. The spatial distribution characteristics and ground-level estimation of $NO_2$ and $SO_2$ over Huaihe River Basin and Shanghai based on satellite observations, *Proc. of SPIE 10767, Remote Sensing and Modeling of Ecosystems for ustainability 54*, 107670L SPIE optical Engineening, California, USA, Sep. 18, p. 10767.

Zhang, H., B. Di, D. Liu, J. Li, and Y. Zhan, 2019. Spatiotemporal distributions of ambient $SO_2$ across China based on satellite retrievals and ground observations: Substantial decrease in human exposure during 2013-2016, *Environmental Research*, 179: 108795.